Predicting Protein Levels In Cells

Marco Kemmerling
Supervisor: Rachel Cavill

Maastricht University, 22.06.16

Overview

- Data from TCGA network (http://cancergenome.nih.gov)
- Linear Regression
- Support vector regression, kernels
- Feature selection
- Gene data
- Datasets

Linear Regression (Rawlings, 1998)

- Fit a linear model to the data
- $y = X\beta + \epsilon$
- Minimise residual sum of squares between predicted and observed dependent variable

Support Vector Regression (Smola, 2004)

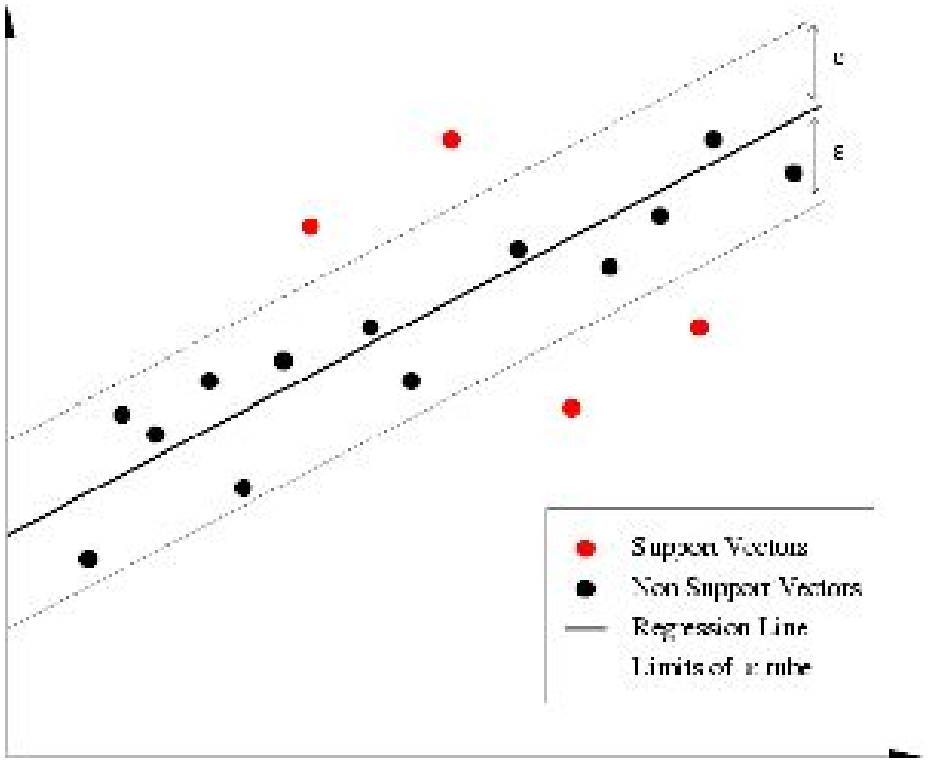


Figure 1: SVR (Gilardi, 2001)

Kernel trick

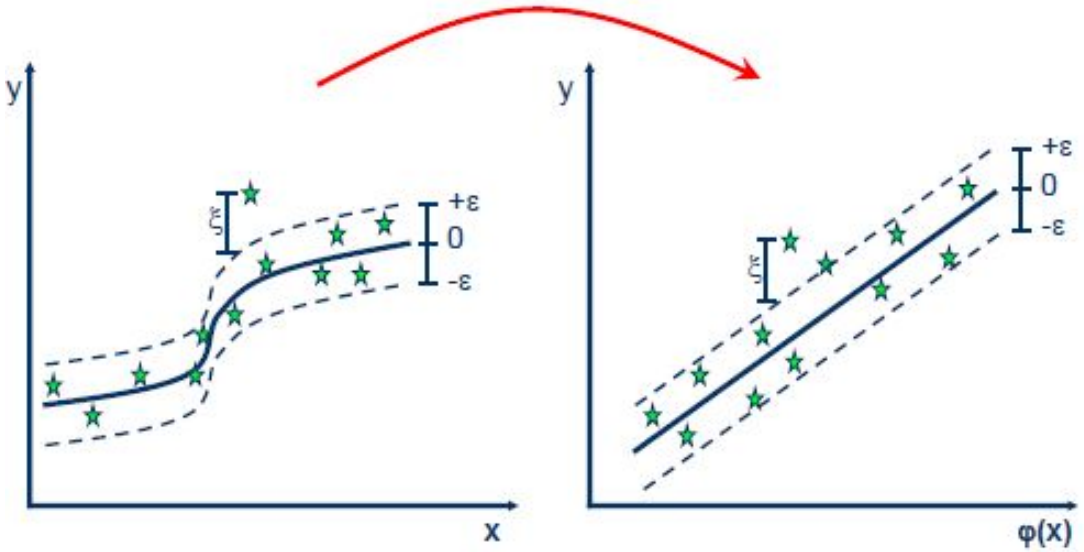
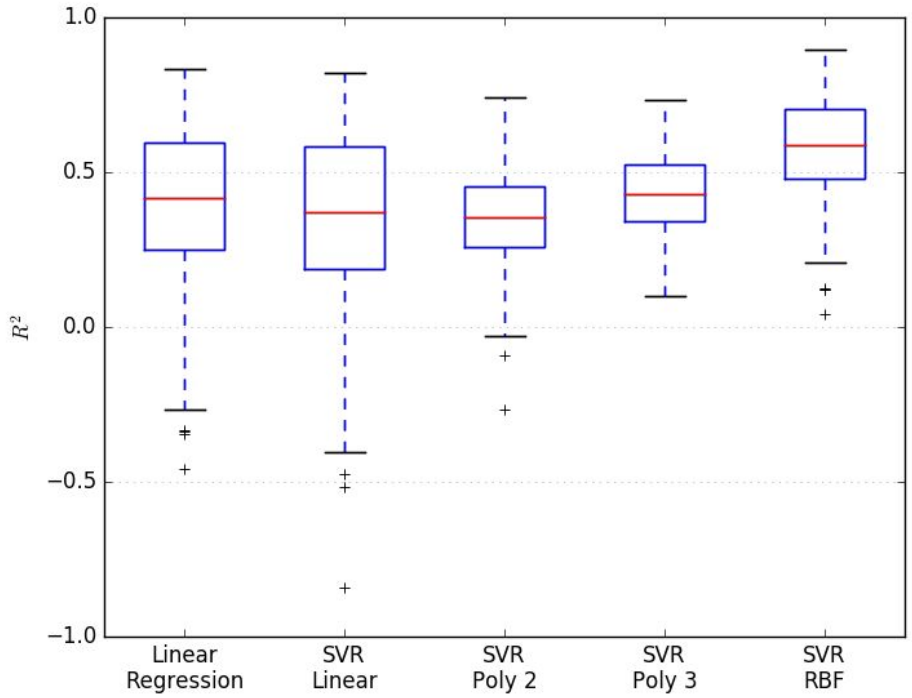


Figure 2: Kernel trick (Sayad, 2016)

Kernels:

- Linear
- Polynomial
- Radial basis function (RBF)

Results: LR vs SVR



Method	Mean R^2
Lin. Reg.	0.398
SVR Linear	0.346
SVR Poly 2nd	0.354
SVR Poly 3rd	0.430
SVR RBF	0.577

Feature selection

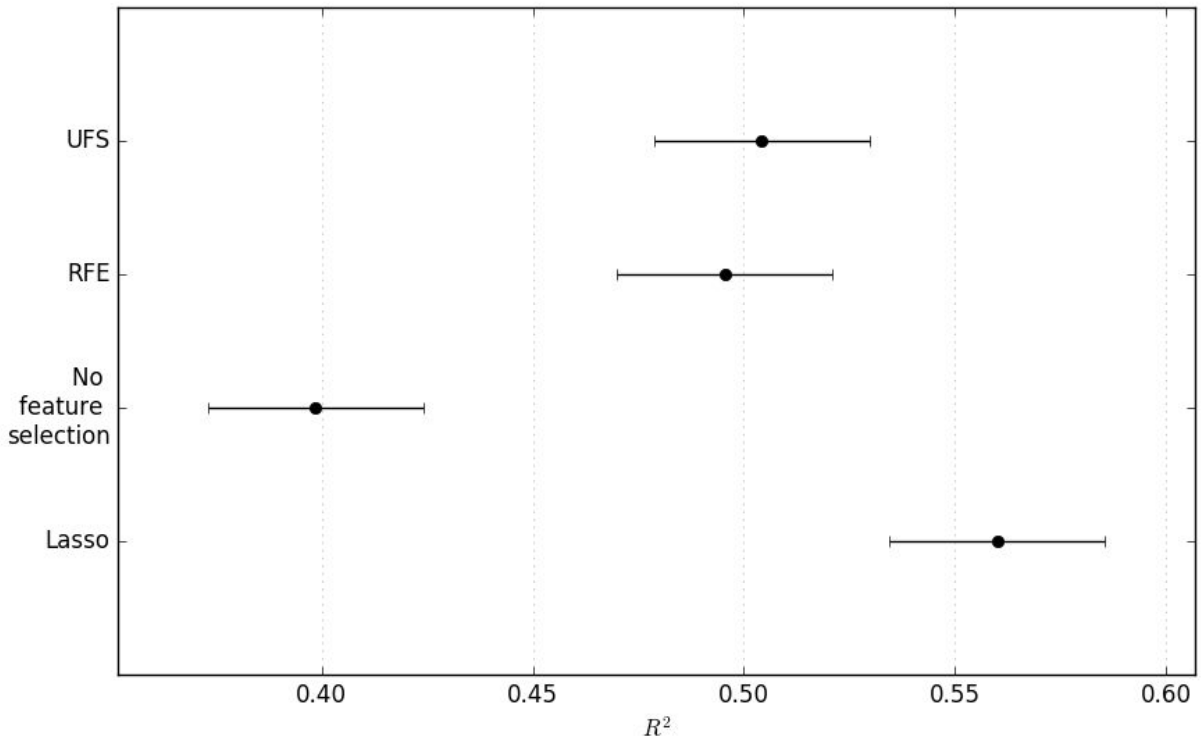
- Univariate feature selection (UFS)

  → f-test (Zhang, 1997)

- Recursive feature elimination (RFE) (Guyon, 2002)

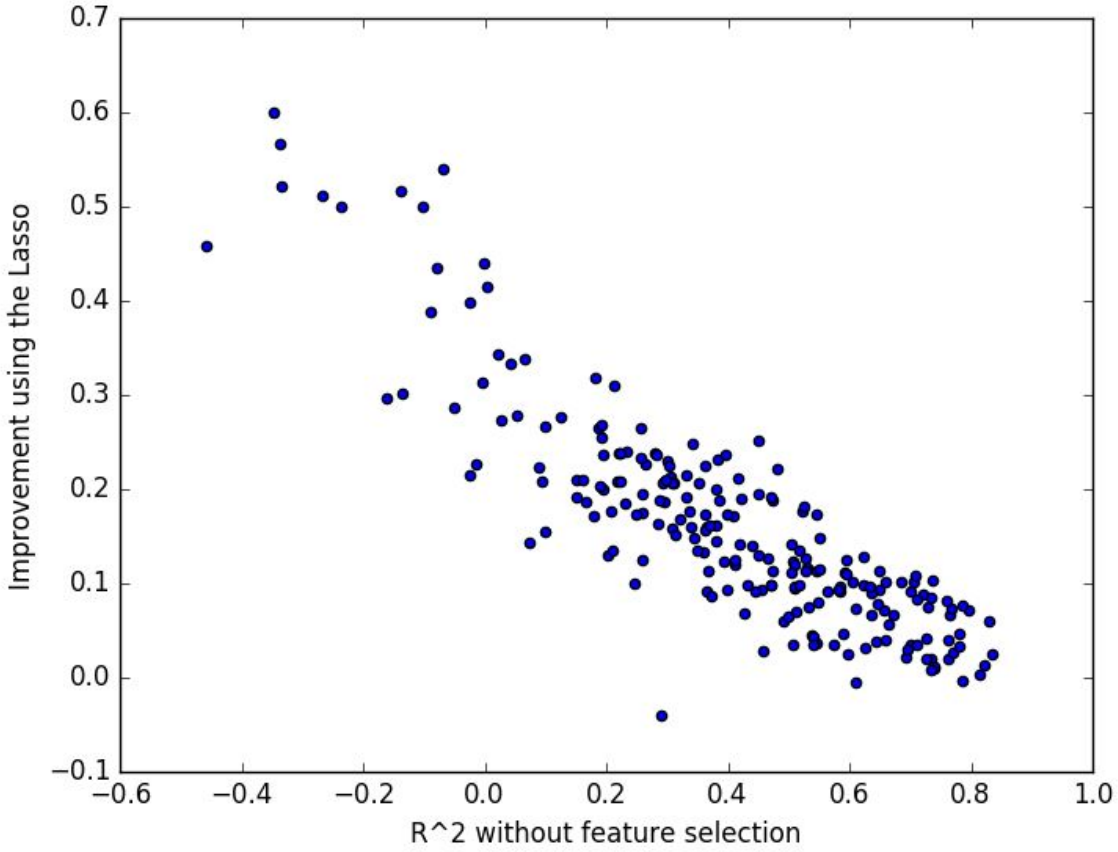
- The Lasso (Tibshirani, 1996)

Results: feature selection

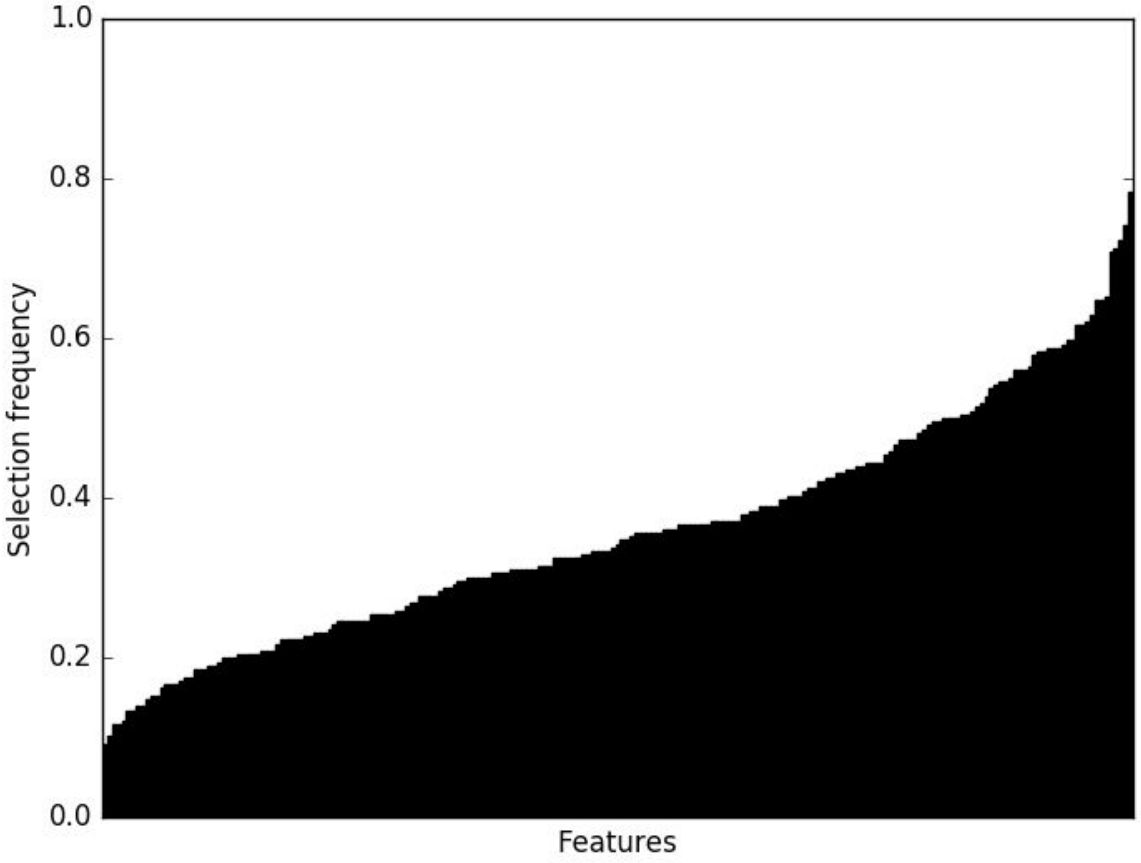


	SVR RBF	LR with Lasso
Mean R^2	0.577	0.560

Which proteins benefit the most?



Are some proteins generally more useful?



Including gene data

- Including all available gene data overwhelms algorithms
- **Approach A**: Proteins have one or more genes which code for them

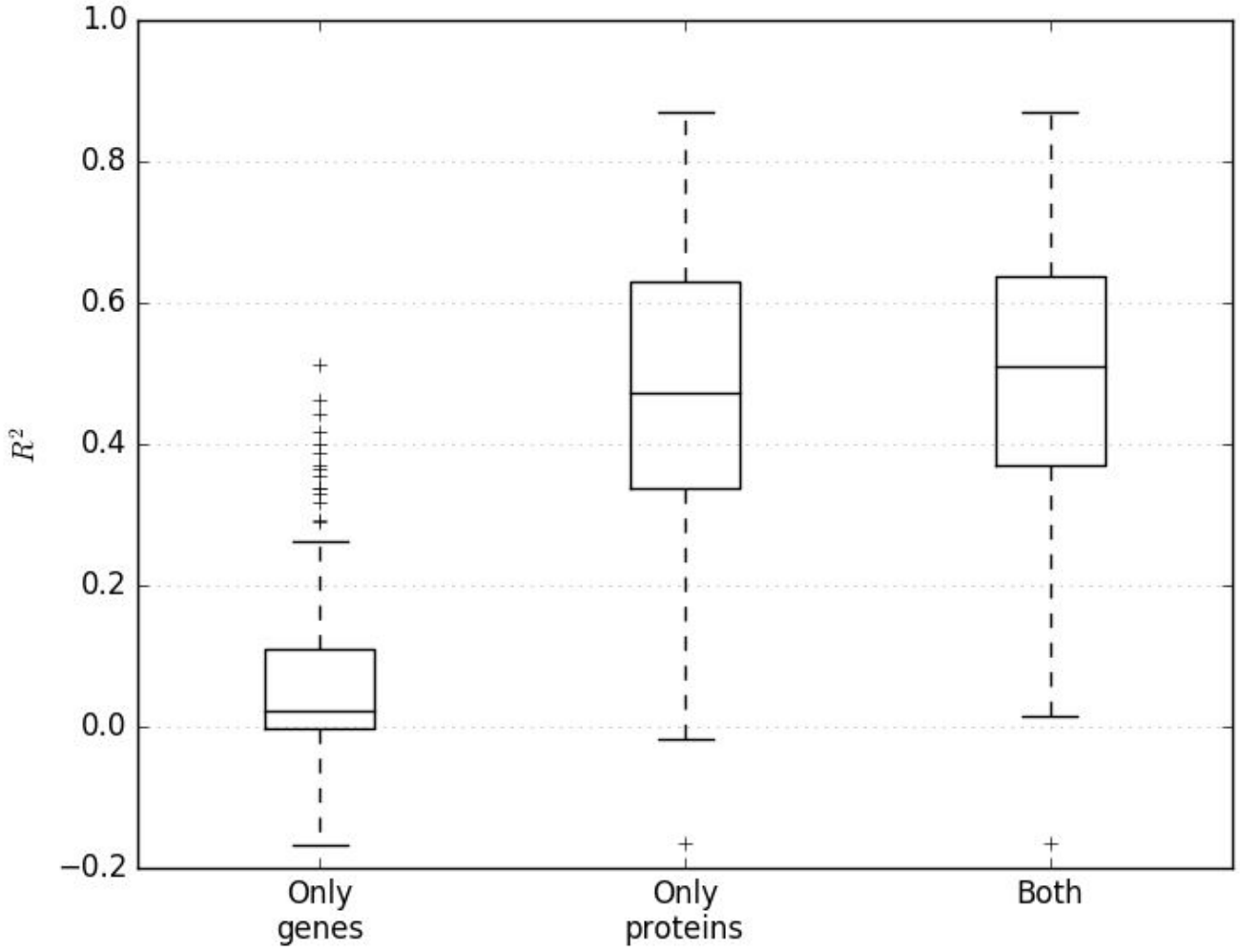
  → Use only coding genes

- **Approach B**: Genes can be grouped in pathways, in which genes interact

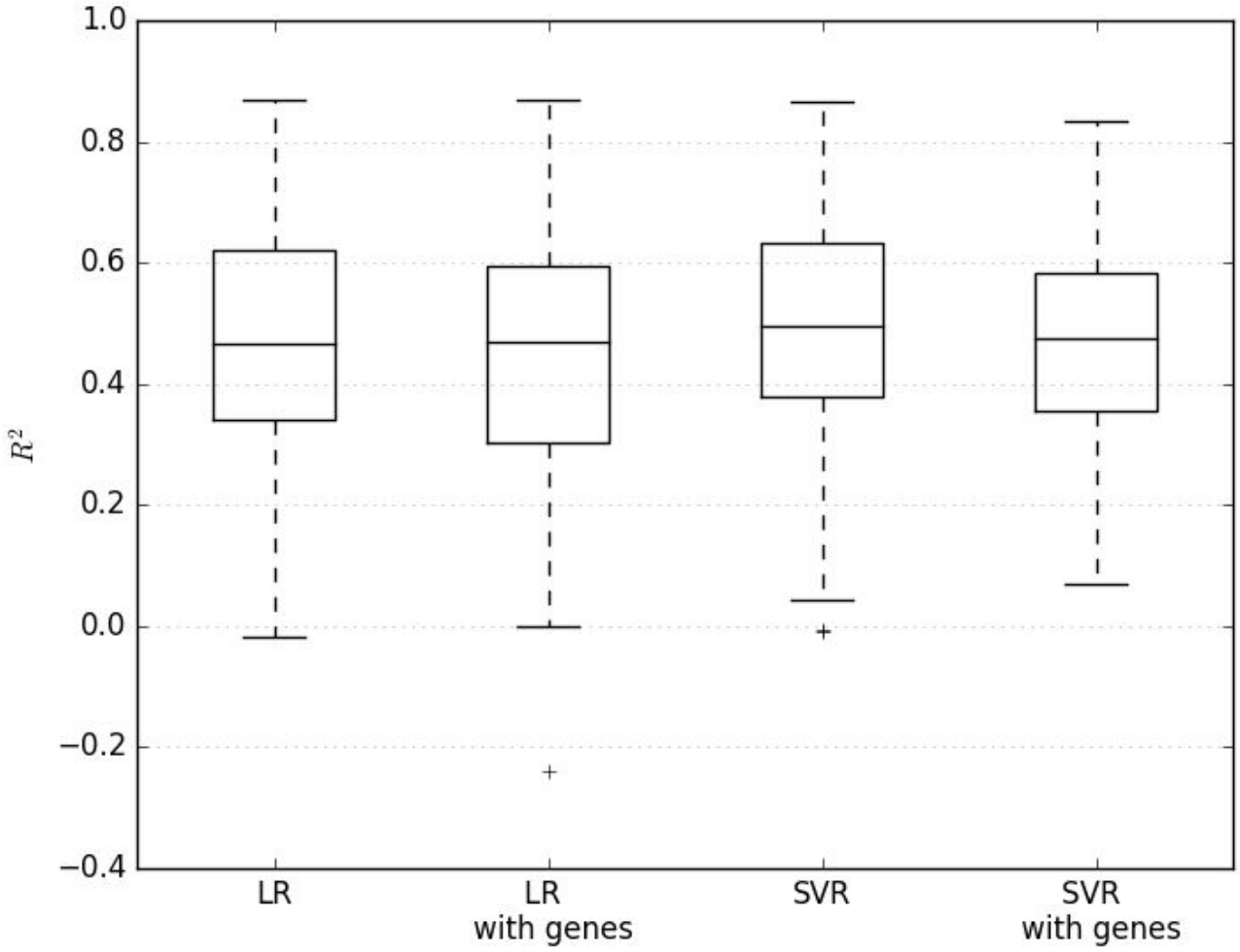
  →Use all genes from all pathways which include

  coding gene

Approach A: Coding genes (SVR)



Approach B: Pathway genes



Conclusion

- SVR with RBF kernel most useful
- Feature selection improves Linear regression performance
- Gene data not very useful.

Reference list

- Rawlings, J. O., Pantula, S. G. & Dickey, D. A. (1998). Applied Regression Analysis: A Research Tool. Springer.
- Smola, A. J. (2004). A tutorial on support vector regression. Statistics and Computing.
- Gilardi, N. & Bengio, S. (2001). File:fig_svr_small.gif. [Image file]. Retrieved from: http://publish.uwo.ca/~jmalczew/gida_7/Gilardi/fig_svr_small.gif
- Sayad, S. (2016). File:SVR_5.png. [Image file]. Retrieved from: http://www.saedsayad.com/images/SVR_5.png
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Wiley.
- Zhang, J. (1997). Selecting the best regression equation via the P-value of F-test. Metrika.
- Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning.